Approaches to Image Processing for Land Use and Land Cover Classification (2^{nd} DRAFT)

Prepared for the project "Remote Sensing Applications for Environmental Analysis in Transportation Planning: Application to the Washington State I-405 Corridor"

Demin Xiong, Russell Lee and J. Bo Saulsbury
Oak Ridge National Laboratory

Elizabeth Lanzer and Albert Perez
Washington State Department of Transportation

June 2002

1. Introduction

A major task of the ongoing project is to establish approaches to derive Land Use and Land Cover (LULC) classification information using remotely sensed imagery and spatial data in existing Geographic Information Systems (GIS). The purpose is to demonstrate the applicability of commercial remote sensing products and spatial information technologies to environmental analysis in transportation planning, using the I-405 corridor in Washington State as a test case. This document focuses on the proposed technical approaches that are to be used for image processing in LULC classification.

Several technical strategies are considered in formulating the proposed approaches. These strategies include (1) the use of a supervised classification scheme; (2) the use of multi-spectral, multi-resolution and multi-source imagery; and (3) the integration of existing GIS data and remotely sensed data. Each of these strategies is briefly described below.

As one of the major strategies, a supervised classification method will be utilized for LULC classification. Supervised classification involves an image analyst identifying spectral data groupings as a specific information classes. The major advantage of this method is that known LULC classification in some areas can be utilized to derive LULC information in places where LULC classification is unknown. Image processing with a supervised classification method usually starts with selecting training samples. After these training samples are selected, image characteristics, such as spectral intensity statistics, and shapes and patterns of given LULC classes, will be extracted. The extracted image characteristics are also called image signatures because they uniquely identify different types of classes on the ground. By using these signatures, LULC classes can then be identified in the entire study area. In general, ground truth data are needed both in selecting training samples and in validating classification results. Therefore, field trips to collect ground truth data are a crucial part of image data analysis.

The second important strategy is the use of multi-spectral, multi-resolution and multi-source imagery. Even when special care is taken in image data selection, an image with a single vantage point, a given spectral region, and a fixed spatial resolution usually has its limitations. In contrast, the combined use of a variety of image data sources can achieve synergistic results. The use of multi-source imagery will take advantage of several existing technical tools in the commercial image processing system (i.e., ERDAS IMAGINE), which has been selected for the current project. These tools include image-reprojection (which brings images with different coordinate systems into a unified coordinate system), geometric rectification (which provides spatial adjustments such as rubber sheeting to allow precise spatial correspondences among different layers of images), and image sharpening or resolution merging (which is particularly useful when images come with different resolutions).

The third strategy is to make use of existing GIS data. Existing GIS data, such as road networks, hydro-networks and administrative boundaries, can be utilized not only as a general reference for a kind of "ground truth" (e.g., the name of a river), but also as a correlation and/or comparison to the remotely sensed data. More importantly, some of the existing data can be directly utilized in the LULC classification process. This is particularly valuable in situations where image spectral information is not sufficient in identifying different LULC classes. For instance, it is usually difficult to differentiate between forestland and densely-wooded residential areas when spectral information alone is utilized. But if Census population data can be referenced, then the differences between forestland and woody residential areas can be drawn based upon whether residential houses are present in the given area.

The rest of this document is organized as follows. Section 2 provides a rationale for the use of the LULC classification scheme. Section 3 describes the planned collection of ground truth data. Section 4 discusses the image and GIS data to be used for the LULC analysis. Section 5 provides technical details of the image processing and data analysis for LULC classification.

2. Land Use and Land Cover Classes

Different LULC classification systems have been developed to facilitate the documentation of LULC information. The USGS LULC classification by Anderson et al. (1976) is one of the systems that has been widely adopted in the remote sensing and GIS community because it was designed with a consideration of the use of remotely sensed data. The USGS LULC classification is a hierarchical system that has different levels of classification. Usually only the top two levels of classification (e.g., level 1 and level 2) are needed for a given application. The top level (level 1) classification consists of nine categories: 1-Urban or built-up land, 2-Agricultural land, 3-Rangeland, 4-Forest land, 5-Water, 6-Wetland, 7-Barren Land, 8-Tundra, and 9-Perennial snow or ice. Each category at the top level is further divided into subcategories (e.g., Urban or built-up land has seven subcategories, including: 11-Residential, 12-Commercial or services, 13-Industrial, 14-Transportation, communication, utilities, 16-Mixed urban or built-up land, and 17-Other urban or built-up land).

Actually, the USGS classification scheme is not an exact match of the LULC categories as commonly used in Environmental Impact Statement (EIS) (Saulsbury, 2002). In particular, there are some LULC categories that are used for the EIS purposes that are not represented in the

USGS LULC classification. For instance, wildlife habitat and threatened and endangered species are identified as required LULC classes for most EIS work, but are absent in the USGS classification system. The same is true for a couple of other LULC categories (e.g., floodplains, recreational resources and historic, cultural, and archaeological resources). Nevertheless, the USGS classification system has a reasonable compatibility with the LULC categories required in an EIS, and many classes required for EIS purposes can be mapped from USGS LULC classes. From a technical standpoint, the USGS classification is amenable to the remotely sensed data. That is, the LULC classification is compatible with the discerning capabilities of the remotely sensed data.

Given the differences between EIS requirements and USGS LULC classes, the proposed strategy is to use the USGS classification as a starting point. With this classification scheme, major LULC classes will first be identified with the remotely sensed data. In a later stage, when additional data such as utility data, census population data or habitat data are utilized or become available, USGS LULC classes will then be realigned into EIS categories.

3. Ground Truth and Field Trips

To select training samples for LULC classification and to validate LULC classification results, the collection of ground truth data through field trips is an integral part of the LULC classification process. The project plans two field trips: the first will be conducted before the LULC classification process and the second will be conducted at the end of the process. The purpose of the first field trip is to gain familiarity with the study area and to obtain training samples for the LULC classes. The purpose of the second field trip is to provide validation, and verification, and modification to the classification results.

In a practical sense, the collection of the ground truth data is not limited to the field trip and can begin before the field trip. Through the study of remotely sensed imagery, existing GIS data and data from other sources, an overall picture of the study area in terms of LULC can be obtained. When information from different sources correlates highly, this information becomes highly reliable and can then be treated as the ground truth. Actually, many training samples of LULC categories can be directly obtained through the use of this type of information, such as water bodies, forestland, and residential areas. Ground truth obtained through in-house data analysis can significantly reduce the workload of field trips and can also help in planning the field study.

An important part in planning the field trip is the selection of the sites to be visited and the data to be collected. Figure 1 is a map of the sites selected for the first field trip. To help with the planning process, images and existing GIS data are first analyzed to determine known and unknown LULC categories and their locations in the study areas. For the sites where LULC classes are obvious, field observations on the ground will be eliminated or only a quick verification check will be necessary. Only those sites where the LULC classes cannot be determined, or where they might provide representative spectral signatures for the consequent classification, will be planned for a field visit. The data collected for a given site can be more specifically defined with existing information. For instance, when LULC data are collected for a site in a forest area with limited households, the data collection on that site can be limited to whether the site belongs to either forestland or residential area.

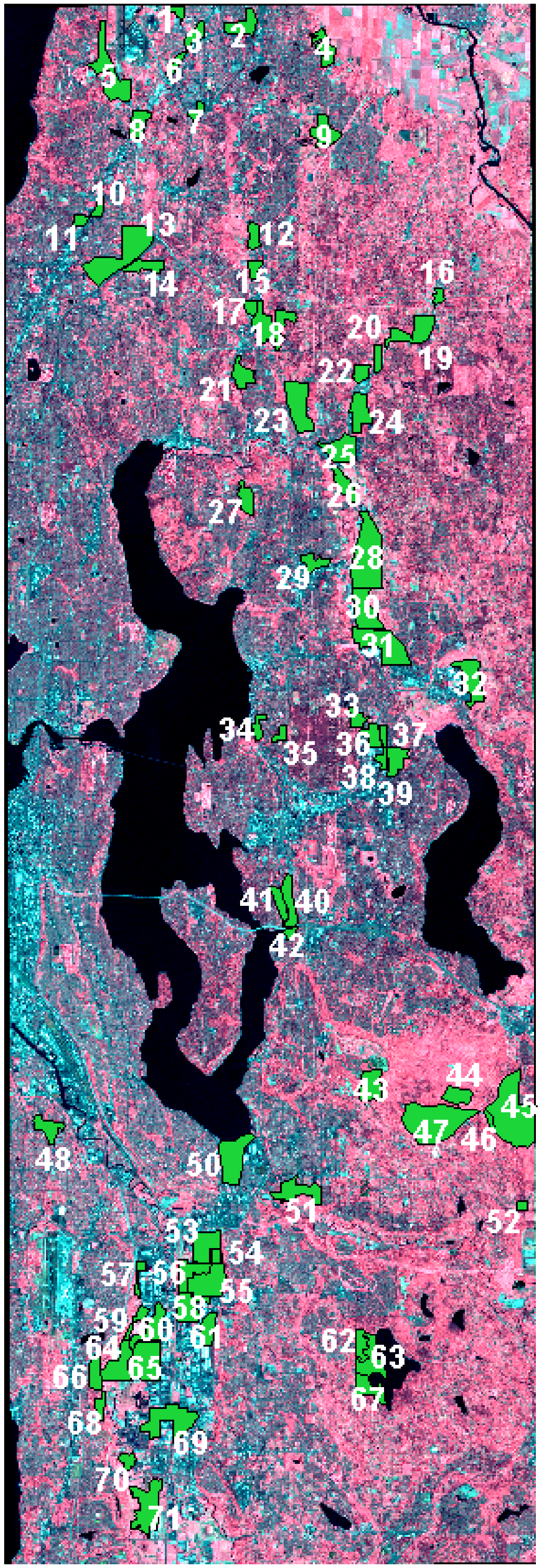

Figure 1. Sites to Be Visited for the First Field Trip.

LULC classes on the ground are changing over time, and some of the sites represent a transitional state of the LULC (e.g., forestland is cleared for residential development). An understanding of the general LULC trend in the study area through interviews or discussions with local government or planning personnel is one of the tasks in the process of ground truth data collection. This kind of understanding will be especially helpful in dealing with data inconsistencies when these data are collected at different times and when LULC classes change.

The second field trip will occur after image analysis results are developed. The major purpose of this field trip is to verify and validate LULC classification results. Similar to the first field trip, a great portion of the verification and validation task will first be carried out in-house. That is, criteria that are used to measure LULC mis-classification or mis-delineation will be first established. Then sample sites will be selected in order to check the classification results. For those sample sites where reliable ground information can be obtained from existing sources (e.g., images or GIS data), existing data will be utilized as the ground truth. The field trip then is to check those sites where LULC classes cannot be determined in-house.

4. Imagery and GIS Data

A key strategy of the project is to use a variety of image data, including black-and-white orthophotos, color orthophotos, and Landsat-7's Enhanced Thematic Mapper Plus (ETM+) to take advantage of their different attributes to achieve synergistic results. For instance, the black-and-white orthophotos and color orthophotos have a one-meter or higher spatial resolution, which can be utilized to achieve high geometric accuracy for ground feature identification and delineation. Landsat ETM+, on the other hand, will be utilized to cover large geographic areas, and provide multi-spectral information that is particularly useful for discriminating among different LULC types.

Landsat ETM+ data have been chosen as the major data source for LULC analysis. This selection is based on consideration of several factors. The I405 EIS is a programmatic EIS, not a site specific EIS, so the requirements for spatial detail can be met by Landsat resolution. The Landsat data are inexpensive and provide large geographic coverage. The data are also collected constantly and continuously for most parts on the Earth. More importantly, they provide seven spectral bands between 0.5 to 12.6 μm with a resolution of 30 x 30 meters (60 x 60 meters for band 6) plus a panchromatic band with a resolution of 15 x 15 meters. To complement the Landsat data, color digital orthophotos and black-and-white orthophotos can provide further details on local land use patterns. For this purpose, ½-foot color digital orthophotos, and 1-meter black-and-white orthophotos have also been collected along the corridor.

The project will also use some of the IKONOS 1-m panchromatic and 4-m multi-spectral images for experimentation and comparison purposes. SpaceImaging contributed these images. Noticeably, the available IKONOS data do not directly cover any part of the study area. Nevertheless, these data will provide a useful reference when comparison between the Landsat data and IKONOS data is necessary.

Existing GIS data are also considered vital data sources for the proposed LULC analysis. These data include political boundaries, hydrology, transportation networks, Census population data, existing land use maps and several other data layers. These data provide valuable information about social, economic and environmental conditions of the study area that might not directly or quickly obtained from the images.

5. Image Processing and LULC Classification

To effectively manipulate and analyze imagery and GIS data, the IMAGINE image processing system has been selected to perform major image and data processing tasks. IMAGINE, which is a commercial product by ERDAS, provides a comprehensive set of functions for image mapping and visualization, image processing and advanced remote sensing. Major image processing tasks are designed to take advantage of the existing functions of the software.

5.1. Data Integration

Because of the use of various types of data, data integration becomes the first task of data processing. At the very beginning, format conversion is necessary to import data that come with different data formats. For instance, Landsat ETM+ data and IKONOS data come with the GeoTIFF format. Census data come as ASCII files. To allow effective data analysis, two types of data formats are used as the standard data formats for the project: the IMAGINE img file format is used for image data and ESRI's shapefile and coverage are used for vector data layers. Data that come in the Mr.Sid file format (e.g., digital orthophotos) are kept with the format since these data come in huge volumes, and compression as provided by the Mr.Sid data format is necessary to make the data display and manipulation tasks tractable.

One of the simplest requirements for data integration is to allow data from different sources to be referenced in the same spatial domain so that these data can be effectively overlaid with each other. Because different data sources make use of different map projections, converting the project data into a standard map projection becomes necessary. The project has chosen State Plane Coordinates (North Zone meters, NAD83) as the standard map projection for the study area. Noticeably, data inaccuracy can also result in spatial inconsistency among different data sources even though the same map projection is utilized.

To handle situations where images come with different resolutions, an image sharpening procedure will be utilized to interpolate multi-spectral, lower resolution imagery such as Landsat-7 ETM+ multi-spectral onto panchromatic high-resolution imagery such as the ETM+ panchromatic band. This is done so that the LULC classification will be performed on the images that presumably have an improved spatial resolution. Several alternative approaches can be used for image sharpening, such as Principal Component analysis, Multiplicative Algorithm, and Brovey Tranvey Transform, which are all standard functions in the IMAGINE software. The Principal Components Analysis is a sophisticated method that is capable of closely maintaining the original scene radiometry after sharpening. The drawback of the method is its computational performance, which, however, appears to be acceptable in the current case, as an initial experiment has indicated. Therefore, the Principal Components Analysis algorithm will be utilized for the image sharpening process.

5.2. LULC Classification

After the first field trip, a set of training samples, each representing a specific type of land use, will be established. Once the locations of the training samples are pinpointed on the map, their image characteristics such as spectral intensity statistics can be identified and extracted from the image. Consequently, the image characteristics are then utilized to classify the entire image to obtain LULC classes for the study area. This standard supervised classification method provides a general framework for LULC classification. However, due to complex spatial patterns of LULC classes in the study area, spectral signatures given by remotely sensed data alone are insufficient in identifying some of the LULC classes.

The I-405 corridor is primarily an area with a mixture of urban and suburban landscapes. Much of the land along the corridor can be identified as urban built up using the USGS classification scheme. Yet, a great portion of the land is covered with trees and grasses, which can be easily confused up with agricultural land or forestland. These are key distinctions for the EIS as they are indicators of impervious surface coverage, which is a major limiting factor to transportation system development. For this reason, some enhancement strategies are needed in order to improve the classification results when the standard supervised classification procedure is utilized. One of the proposed strategies is to combine image data and non-image data during the LULC classification process. Image data usually contain rich spectral information about the ground and can be updated rapidly, but they also present data that can be difficult to interpret effectively in an automated environment. In contrast, existing data provide information that is already validated or verified and in some cases they may provide information that can not be derived directly from an image (e.g., socio-economic activities).

Two types of nationally available, non-image data have been found particularly useful for the enhancement of LULC classification: the USGS LULC maps and the 2000 Census population data. The USGS LULC maps were created at different dates between mid 1970s to early 1980s. One idea of using the USGS LULC maps is to take the LULC patterns represented on the USGS maps as the starting point of the current LULC classification. With this starting point, regularity of LULC conversion of different LULC classes will be analyzed and conversion rules will be defined. Therefore, current LULC classes then can be identified not only with their image characteristics, but also the given USGS LULC patterns and the potential changes that have taken place during the past. The use of the USGS LULC data is based upon the assumption that trends can be identified for the LULC conversion process. That is, LULC changes tend to have certain kinds of regularity. For example, the forestland in the study area will likely be converted into urban land, or possibly agricultural land. At the same time, the likelihood of conversion of build-up areas to agricultural or forestland will be small.

Population data are useful in several circumstances. In urban areas, highly concentrated residential, commercial and industrial land use may have similar image characteristics, which makes it difficult to identify their differences. The use of population and household counts in these areas can provide additional evidence on whether the areas under question have the presence of residential housing. In suburban areas, low-density housing with extensive coverage of trees and grasses can be easily confused with forest and agricultural land. The population data can also be very helpful in resolving these differences because the presence of houses will be a clear indication of residential land use.

Sophisticated tools will be required to make good classification decisions when various kinds of non-image data are utilized. The project plans to use the tool called IMAGINE Expert Classifier that comes with IMAGINE software. The IMAGINE Expert Classifier has two modules: the Knowledge Engineer and the Knowledge Classifier. The Knowledge Engineer can be used to construct a knowledge base that defines input variables, decision rules and output classes. The Knowledge Classifier provides tools to apply an existing knowledge base and to generate classification.

Texture analysis is another strategy that has been considered for LULC classification enhancement. It is well known that in some cases, image intensity or spectral information on individual pixels cannot be effectively utilized to establish LULC classes. This is particularly true in urban areas and suburban areas where LULC classes show complex patterns or when different LULC classes have similar spectral characteristics (e.g., forest versus a fruit farm). The texture analysis looks into textures or patterns formed from a group of pixels in a neighborhood area. To do so, recurring spatial characteristics can be captured and utilized to distinguish closely resembling LULC classes.

A variety of texture measurements can be generated as indices for texture analysis, which include maximum probability, moments, contrast, homogeneity, entropy, etc. The project will focus on a couple of texture measurements that can be directly generated with the ERDAS IMAGINE software. Given different types of texture measurements and the possible sizes of neighborhood areas that can be chosen, experiments will be conducted to select some good texture measurements that can effectively improve the classification results. It is worth noting that texture measurements will be generated on each image pixel. Therefore, these measurements will be treated as additional image bands to be used directly in the LULC classification process.

5.3. Additional Data Analysis

The image processing and LULC classification will produce a LULC map for the study area. This map will initially show the USGS LULC classes. Additional data analysis will include realigning the USGS classification into EIS classification categories and linking LULC maps with transportation infrastructure and its potential impact to the environment.

To align USGS LULC classes into EIS categories, data on wildlife habitat, utilities, historical, cultural and archaeological resources will be used. These data, when available, will be overlaid with the LULC map to generate LULC categories that are suitable for EIS purposes. To link the LULC with the transportation infrastructure and to demonstrate the effect of potential transportation development and activities on the environment, drainage areas in along the corridor will be with the LULC map so as to produce statistics of LULC resources that may be potentially impacted along the corridor. Details regarding data analysis will be given in a separate document on approaches to data analysis.